A quick glance at the market situation

Both Hadoop and Spark are open source projects by Apache Software Foundation and both are the flagship products in big data analytics. Hadoop has been leading the big data market for more than 5 years. According to our recent market research, Hadoop's installed base amounts to 50,000+ customers, while Spark boasts 10,000+ installations only. However, Spark's popularity skyrocketed in 2013 to overcome Hadoop in only a year. A new installation growth rate (2016/2017) shows that the trend is still ongoing. Spark is outperforming Hadoop with 47% vs. 14% correspondingly.

The key difference between Hadoop MapReduce and Spark

To make the comparison fair, here we will contrast Spark with Hadoop MapReduce, as both are responsible for data processing. In fact, the key difference between them lies in the approach to processing: Spark can do it in-memory, while Hadoop MapReduce has to read from and write to a disk. As a result, the speed of processing differs significantly – Spark may be up to 100 times faster. However, the volume of data processed also differs: Hadoop MapReduce is able to work with far larger data sets than Spark.

Now, let's take a closer look at the tasks each framework is good for.

Tasks Hadoop MapReduce is good for:

- **Linear processing of huge data sets.** Hadoop MapReduce allows parallel processing of huge amounts of data. It breaks a large chunk into smaller ones to be processed separately on different data nodes and automatically gathers the results across the multiple nodes to return a single result. In case the resulting dataset is larger than available RAM, Hadoop MapReduce may outperform Spark.
- **Economical solution, if no immediate results are expected.** Our Hadoop team considers MapReduce a good solution if the speed of processing is not critical. For instance, if data processing can be done during night hours, it makes sense to consider using Hadoop MapReduce.

Tasks Spark is good for:

- **Fast data processing.** In-memory processing makes Spark faster than Hadoop MapReduce – up to 100 times for data in RAM and up to 10 times for data in storage.
- **Iterative processing.** If the task is to process data again and again – Spark defeats Hadoop MapReduce. Spark's Resilient Distributed Datasets (RDDs) enable multiple map operations in memory, while Hadoop MapReduce has to write interim results to a disk.
- **Near real-time processing.** If a business needs immediate insights, then they should opt for Spark and its in-memory processing.
- **Graph processing.** Spark's computational model is good for iterative computations that are typical in graph processing. And Apache Spark has GraphX – an API for graph computation.
- **Machine learning**. Spark has MLlib – a built-in machine learning library, while Hadoop needs a third-party to provide it. MLlib has out-of-the-box algorithms that also run in memory. But if required, our Spark specialists will tune and adjust them to tailor to your needs.
- **Joining datasets.** Due to its speed, Spark can create all combinations faster, though Hadoop may be better if joining of very large data sets that requires a lot of shuffling and sorting is needed.

Examples of practical applications

We analyzed several examples of practical applications and made a conclusion that Spark is likely to outperform MapReduce in all applications below, thanks to fast or even near real-time processing. Let's look at the examples.

- **Customer segmentation.** Analyzing customer behavior and identifying segments of customers that demonstrate similar behavior patterns will help businesses to understand customer preferences and create a unique customer experience.
- **Risk management**. Forecasting different possible scenarios can help managers to make right decisions by choosing non-risky options.
- **Real-time fraud detection.** After the system is trained on historical data with the help of machine-learning algorithms, it can use these findings to identify or predict an anomaly in real time that may signal of a possible fraud.
- **Industrial big data analysis.** It's also about detecting and predicting anomalies, but in this case, these anomalies are related to machinery breakdowns. A properly configured system collects

the data from sensors to detect pre-failure conditions.

Which framework to choose?

It's your particular business needs that should determine the choice of a framework. ***Linear processing of huge datasets*** is the advantage of Hadoop MapReduce, while Spark delivers ***fast performance***, ***iterative processing, real-time analytics, graph processing, machine learning and more***. In many cases Spark may outperform Hadoop MapReduce. The great news is the Spark is fully compatible with the Hadoop eco-system and works smoothly with Hadoop Distributed File System, Apache Hive, etc.

Big Data Analytics & Consulting

Big data is another step to your business success. We will help you to adopt an advanced approach to big data to unleash its full potential.

BIG DATA SERVICES
By Alex Bekker
Sep 14, 2017
CIO Blog , Big Data , Data Analytics
2 Comments
Read more

Share on LinkedIn

Share on Facebook

Share on Google +